基于文本生成技术的历史古籍事件识别模型构建研究1

王彦莹[1,2]，*王昊[1,2]，朱惠[1,2]，李晓敏[1,2]
[1]（南京大学信息管理学院，南京 210023）
[2]（江苏省数据工程与知识服务重点实验室（南京大学），南京 210093）

摘要：
[目的] 对比序列标注方法和文本生成方法在历史古籍事件识别上的表现，以构建历史古籍事件识别模型。
[方法] 本文选取《三国志》为原始语料，序列标注实验对《三国志》事件数据集进行 BMES 标注，构建 BBCN-SG 模型，文本生成实验构建 T5-SG 模型，对比两种方法的表现。又构建 RoBERTa-SG、NEZHA-SG 模型展开生成模型的对比实验。结合三个文本生成模型，融入 Stacking 集成学习的思想，构建 Stacking-TRN-SG 模型。
[结果] 在历史古籍事件识别建模问题上，文本生成方法的表现明显优于序列标注方法。而在文本生成方法中，三个模型表现则是 RoBERTa-SG ＞ T5-SG ＞ NEZHA-SG。Stacking 集成学习大大提高了生成模型的识别效果。
[局限] 本文计算资源有限，Stacking-TRN-SG 模型缺少在其他历史古籍语料中的应用研究。
[结论] 本文构建的 Stacking-TRN-SG 模型初步实现历史古籍的自动事件识别。

关键字：历史古籍　事件识别　文本生成　序列标注　集成学习
分类号：G254

Research on the construction of event recognition model in historical books based on text generation technology

Wang Yanying[1,2], Wang Hao[1,2], Zhu Hui[1,2], Li Xiaomin[1,2]
[1] (*School of Information Management, Nanjing University, Nanjing 210023, China*)
[2] (*Jiangsu Province Key Laboratory of Data Engineering and Knowledge Services (Nanjing University), Nanjing 210093, China*)

Abstract:
[Objective] In order to construct a event recognition model in historical books, the performance of sequence labeling method in event recognition in historical ancient books is compared with that of text generation method.
[Methods] In this paper, "Three Kingdoms" is selected as the original corpus. To compare the performance of the two methods, performing on the "Three Kingdoms" event data set, the sequence labeling experiment used BMES annotation and builded the BBCN-SG model ,and the text generation experiment builded the T5-SG model.It also builded RoBERTa-SG and NEZHA-SG models to conduct comparative experiments on generative models. Combining three text generation models and integrating the idea of Stacking ensemble learning, the Stacking-TRN-SG model is constructed.
[Results] On the subject of modeling event recognition in historical ancient books, the performance of the text generation method is significantly better than that of the sequence labeling method. In

---

1　通讯作者：王昊，E-mail：ywhaowang@ nju.edu.cn
本文系国家自然科学基金面上项目“关联数据驱动下我国非遗文本的语义解析与人文计算研究”（72074108）和中央高校基本科研项目“面向人文计算的方志文本的语义分析和知识图谱研究”（010814370113）的研究成果之一，并受江苏青年社科英才和南京大学仲英青年学者等人才培养计划的支持。

the text generation method, the performance of the three models is RoBERTa-SG > T5-SG > NEZHA-SG. Stacking ensemble learning greatly improves the recognition performance of generation models.
[Limitations] The computational resources of this paper are limited, and the Stacking-TRN-SG model lacks application research in other historical and ancient corpora.
[Conclusions] The Stacking-TRN-SG model constructed in this paper preliminarily realizes the automatic event recognition of historical ancient books.
Keywords: Historical books Event recognition Text generation Sequence labeling Ensemble learning

1 引言

历史古籍是中华文化源远流长、博大精深的重要标志。将历史古籍数字化、应用化可以更好地传承中华文化.基于历史古籍构建知识图谱可以直观地向人们展示历史，从而了解历史。事件识别是构建知识图谱过程中信息抽取的重要一环。然而，针对历史古籍进行事件识别，通常会面临古汉语语义难以理解、单字居多从而不便概括事件等问题。如果采用人工概括历史事件，耗时耗力，且具有较高的主观性，容易受到研究者知识水平的限制，很有可能出现缺失或错误的情况。因此，面向历史古籍的事件自动识别是十分有必要的。

关于历史古籍的事件识别，本文需要解决以下几个问题：对于《三国志》等历史古籍的古汉语，研究者尚且难以理解其含义，机器能否准确识别事件？序列标注和文本生成方法是目前主流的事件识别方法，但是缺少两种方法的具体比较研究，两者在古汉语上的表现分别如何、有无优劣之分？目前序列标注方法大多数是针对命名实体识别的，标注的标签也以短距离的约束为主，而事件识别则是对于长句中的部分字概括，存在较多远距离约束的情况，对于序列标注方法而言，远距离的约束能否被机器较好地学习？对于文本生成方法而言，机器能否生成类古汉语形式的事件？为此，本文希望通过序列标注和文本生成技术对历史古籍开展事件识别研究，以构建历史古籍事件识别模型，实现历史古籍事件识别的自动化。

本文选取《三国志》为原始语料，分别从序列标注与文本生成两个方法展开实验。序列标注实验对《三国志》事件数据集进行 BMES 标注，应用 BERT-BiLSTM-CRF-NER 模型进行训练和预测。文本生成实验则应用 T5 预训练模型进行训练和预测。文本生成方法在《三国志》事件数据集上的表现大大超过序列标注方法，所以又选取 RoBERTa、NEZHA 两个文本生成模型进行事件识别训练。最后，结合三个文本生成模型，融入 Stacking 集成学习的思想，构建了 Stacking-TRN-SG 模型。

2 相关研究

事件识别[1][2][3]是信息抽取的重要组成部分，至今已取得一定的研究成果。目前国内外的研究工作主要将事件识别分为两大类：基于规则的方法和基于统计的方法。基于规则的方法以模式匹配为主要手段，即事先制定字典，然后根据一定的规则和模式将待识别的句子与字典进行匹配，准确率较高，如 Surdeanu 等[4][5]构建了针对开放域的事件抽取系统 FSA。但这个方法对字典的依赖性较大，可移植性差。基于统计的方法将事件识别作为分类问题，主要研究分类器的选择、构建和特征的选择，常用的方法有隐马尔可夫模型 HMM[6]、最大熵模型 MEM[7]、支持向量机 SVM[8]、条件随机场 CRFs[9]等。这个方法相对来说不需要过多的人工，且更为灵活，可移植性高。如 Ahn D.[10]结合 MegaM、Timbl 两种机器学习方法研究事件类型识别和事件元素识别，在 ACE 英文语料上取得了较好的识别效果。李章超等人应用模式匹配法实现《左传》战争句的识别[19][20][21]，这种方法依赖事件触发词表和规则的构建，不利于广泛性事件识别的开展，并且本文研究的历史古籍事件识别具有古汉语单字居多的特点，在句法上与现代汉语不同，不适合基于规则的方法。在基于统计的方法中，CRFs 模型的特征设计更为灵活，被广泛应用与命名实体识别领域。古汉语命名实体识别[11][12][13][14]主要包括人名、地名等实体，这类命名实体识别的约束以短距离为主，少有远距离约束的研究，因此笔者参考命名实体识别的方法，对历史事件进行序列标注，探究序列标注方法对远距离约束的事件识别的有效性。

本文将文本生成方法的事件识别转化为生成式摘要任务。近几年，深度学习技术不断发展，序列到序列模型（Sequence to sequence, Seq2seq）研究[18]取得了极大进步，被广泛应用于自然语言生成领域。Cho 等[15]和 Sutskever 等[16]提出了 Seq2seq 模型，主要结构是编码器（encoder）和解码器（decoder），其基本思想即通过输入序列的全局信息推断出与之相对应的输出序列。Rush 等[17]首次将 Seq2seq 模型应用于生成式摘要，相比之前的生成式方法，Seq2seq 更加接近于人工生成摘要，取得了良好的生成效果。此后，基于 Seq2seq 的生成式摘要模型的相关研究如雨后春笋般涌现，为机器学习领域作出了极大的贡献。

3 数据与方法

3.1 研究框架

结合上述的模型方法，历史古籍选取《三国志》为原始语料，开展事件识别研究，实验的整体研究框架如图 1 所示。具体如下：

首先对比研究序列标注方法与文本生成方法在《三国志》事件数据集上的表现。在序列模型部分，结合序列标注的方法，对已有的事件数据集进行 BMES 标注，基础模型选用 BERT-BiLSTM-CRF-NER 模型，重新训练后得到 BBCN-SG 模型，对测试集进行预测。在生成模型部分，应用 T5 预训练模型对事件数据集进行训练，得到 T5-SG 模型，并对测试集进行预测。结果发现文本生成方法在事件数据集上的表现明显优于序列标注方法。

基于这一结果，又增加了 RoBERTa、NEZHA 两个预训练模型对事件数据集进行生成实验，分别构建 RoBERTa-SG、NEZHA-SG 模型，并对比 3 个文本生成模型在《三国志》事件数据集上的表现。最后，在此基础上，融入集成学习思想，采用 Stacking 方法将三个生成模型融合，构建 Stacking-TRN-SG 模型。

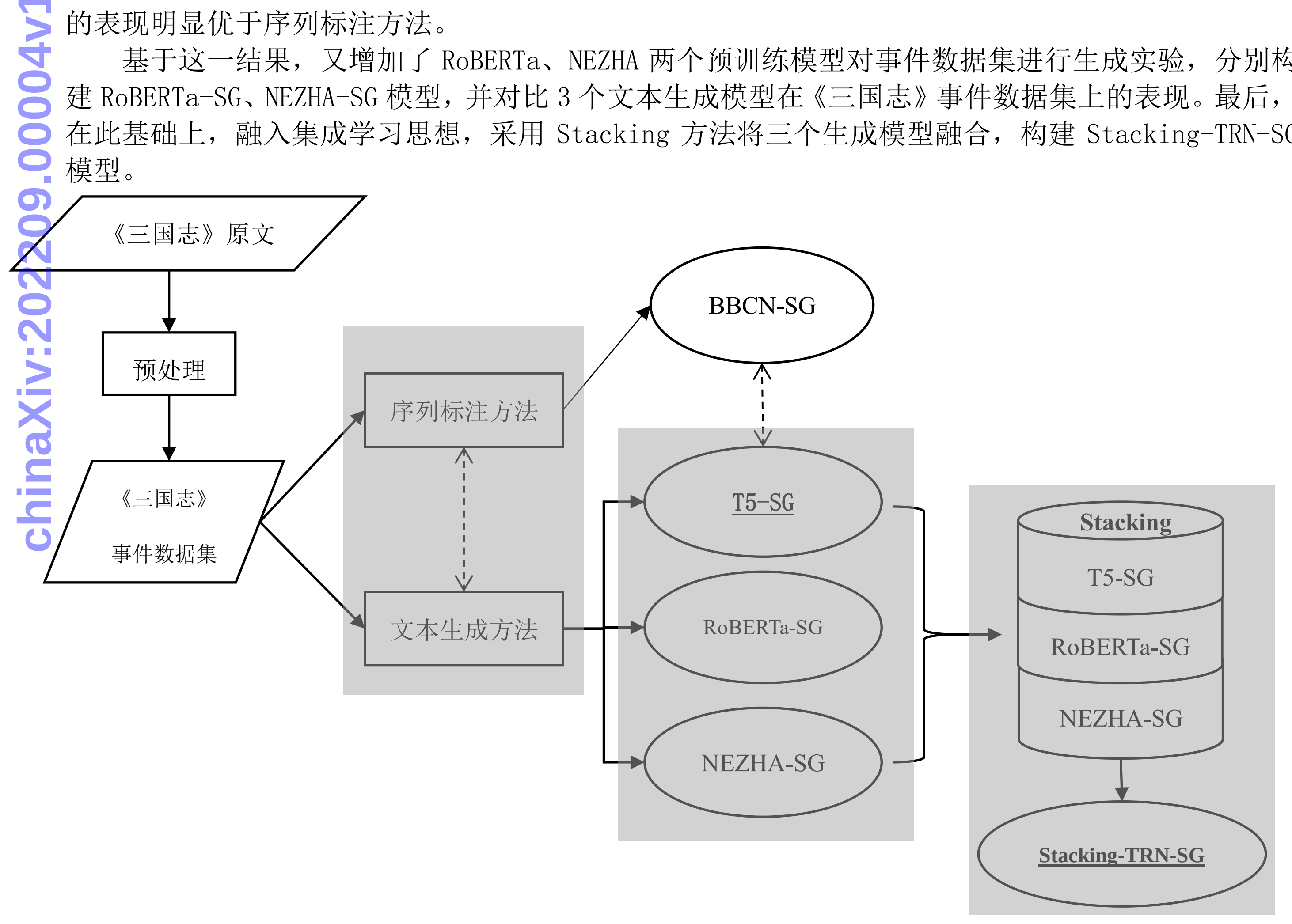


图 1　实验研究框架流程图

3.2 数据收集与预处理

(1) 数据收集

实验数据来源于《三国志》1-30 章，即《魏书》。数据原文是繁体的，以“ . ”为句读，并且在

文中以括号形式加入了注释。将数据原文由繁体转化为简体，将句读由“．”替换为空格，并且删去了括号内的注释。此外，由于历史发展的原因，有部分原文文字无法被机器识别，笔者根据“古诗文网”、“古诗大全”[2]等网站，综合考量，将其进行替换补充。如“士卒无 Z00050 志”替换为“士卒无斗志”。以《三国志》第一章《武帝纪》第一段为例，《三国志》原文的文本处理如表 1 所示，并提供了“古诗文网”的版本作为参考：

表 1　《三国志》第一章《武帝纪》第一段的文本处理

初始版本	太祖武皇帝．沛國譙人也．姓曹．諱操．字孟德．漢相國參之後．（太祖一名吉利．小字阿瞞．王沈魏書曰．其先出於黃帝．當高陽世．陸終之子曰安．是為曹姓．周武王克殷．存先世之後．封曹快於邾．春秋之世．與於盟會．逮至戰國．為楚所滅．子孫分流．或家于沛．漢高祖之起．曹參以功封平陽侯．世襲爵土．絕而復紹．至今適嗣國於容城．）桓帝世．曹騰為中常侍大長秋．封費亭侯．（司馬彪續漢書曰．騰父節．字元偉．素以仁厚稱．鄰人有亡豕者．與節豕相類．詣門認之．節不與爭．後所亡豕自還其家．豕主人大 Z00592．送所認豕．并辭謝節．節笑而受之．由是鄉黨貴歎焉．長子伯興．次子仲興．次子叔興．騰字季興．少除黃門從官．永寧元年．鄧太后詔黃門令選中黃門從官年少溫謹者配皇太子書．騰應其選．太子特親愛騰．飲食賞賜與眾有異．順帝即位．為小黃門．遷至中常侍大長秋．在省闥三十餘年．歷事四帝．未嘗有過．好進達賢能．終無所毀傷．其所稱薦．若陳留虞放．邊韶．南陽延固．張溫．弘農張奐．潁川堂谿典等．皆致位公卿．而不伐其善．蜀郡太守因計吏修敬於騰．益州刺史种暠於函谷關搜得其牋．上太守．并奏騰內臣外交．所不當為．請免官治罪．帝曰．牋自外來．騰書不出．非其罪也．乃寢暠奏．騰不以介意．常稱歎暠．以為暠得事上之節．暠後為司徒．語人曰．今日為公．乃曹常侍恩也．騰之行事．皆此類也．桓帝即位．以騰先帝舊臣．忠孝彰著．封費亭侯．加位特進．太和三年．追尊騰曰高皇帝．）養子嵩嗣．官至太尉．莫能審其生出本末．（續漢書曰．嵩字巨高．質性敦慎．所在忠孝．為司隸校尉．靈帝擢拜大司農．大鴻臚．代崔烈為太尉．黃初元年．追尊嵩曰太皇帝．吳人作曹瞞傳及郭頒世語並云．嵩．夏侯氏之子．夏侯惇之叔父．太祖於惇為從父兄弟．）嵩生太祖．
《古诗文网》版本	太祖武皇帝，沛国谯人也，姓曹，讳操，字孟德，汉相国参之后。桓帝世，曹腾为中常侍大长秋，封费亭侯。养子嵩嗣，官至太尉，莫能审其生出本末。嵩生太祖。
最终版本	太祖武皇帝 沛国谯人也 姓曹 讳操 字孟德 汉相国参之后 桓帝世 曹腾为中常侍大长秋 封费亭侯 养子嵩嗣 官至太尉 莫能审其生出本末 嵩生太祖

(2) **数据预处理**

实验人工提炼出包含事件的原文，并对原文进行缩句摘要。特别地，由于序列模型实验需要，需保证摘要的文字完全来自于原文。处理好的数据如表 2 所示，source 为提炼出的原文文本，target 为 source 的摘要版本，即 target 可以理解为是 source 的子集。最终共得到 671 条数据，数据集以 500：171 的比例划分为训练集和测试集。

表 2　《三国志》事件数据

序号	章节	source	target
1	1	冀州刺史王芬 南阳许攸 沛国周旌等连结豪杰 谋废灵帝 立合肥侯 以告太祖	豪杰谋废灵帝 立合肥侯

[2] “古诗文网”: https://www.gushiwen.cn/guwen/sanguo.aspx

“古诗大全”: https://www.shidaquan.com/ju9656587

2	1	卓到 废帝为弘农王而立献帝	卓废帝为弘农王而立献帝
3	1	卓遂杀太后及弘农王	卓杀太后及弘农王
4	1	太祖至陈留 散家财 合义兵 将以诛卓	太祖将诛卓
5	1	众各数万 推绍为盟主	众推绍为盟主
6	1	到荥阳汴水 遇卓将徐荣 与战不利 士卒死伤甚多 太祖为流矢所中 所乘马被创	太祖遇徐荣 与战不利
7	1	司徒王允与吕布共杀卓	王允与吕布杀卓
8	1	二年春 袭定陶 济阴太守吴资保南城 未拔 会吕布至又击破之 夏 布将薛兰 李封屯巨野 太祖攻之	太祖攻吕布
9	1	卓将李傕 郭汜等杀允攻布	卓将杀允攻布
10	1	公到宛 张绣降 既而悔之 复反 公与战	公与张绣战

生成模型的数据即 source 和 target 数据，而序列模型的数据需要将其转换为标注数据。首先根据 target 数据，对 source 数据进行标注，采用“BMES”标注法，标注实例如表 3 所示，其中具体标注规则如下：① target 数据中的第一个字，在 source 数据上标注为“B”； ② target 数据中的最后一个字，在 source 数据上标注为“E”； ③ target 数据中的其他字，即非首尾字，在 source 数据上标注为“M”； ④ 未在 target 数据中出现的字，在 source 数据上标注为“S”；⑤ target 数据的首尾字，如果多次出现，则都标注为“B”或“E”； ⑥ 标注的句首和句尾的位置，不以在 source 数据中出现的位置为准，而是以在 target 数据中出现的位置为准。根据模型训练的需要，标注数据的每个字及其标注为一行，每条数据间插入空行进行分隔，如表 3 中的实例所示。

表 3 序列标注数据实例

序号		标注实例	实例
1	Source	到荥阳汴水・遇卓将徐荣・与战不利・士卒死伤甚多・太祖为流矢所中・所乘马被创	建 S 安 S 四 S 年 S ・ S **绍 B** 悉 S 军 S **围 M** 之 S ・ S **瓒 E** 遣 S 子 S 求 S 救 S 于 S 黑 S 山 S 贼 S
	Target	太祖遇徐荣・与战不利	
	标注	到 S 荥 S 阳 S 汴 S 水 S・S 遇 M 卓 S 将 S 徐 M 荣 M・S 与 M 战 M 不 M **利 E**・S 士 S 卒 S 死 S 伤 S 甚 S 多 S・S **太 B** 祖 M 为 S 流 S 矢 S 所 S 中 S・S 所 S 乘 S 马 S 被 S 创 S	
2	Source	尚果循西山来・临滏水为营・夜遣兵犯围・公逆击破走之	
	Target	公逆击破走尚	
	标注	**尚 E** 果 S 循 S 西 S 山 S 来 S・S 临 S 滏 S 水 S 为 S 营 S・S 夜 S 遣 S 兵 S 犯 S 围 S・S **公 B** 逆 M 击 M 破 M 走 M 之 S	
3	Source	十二月・孙权为备攻合肥・公自江陵征备・至巴丘・遣张憙救合肥・权闻憙至・乃走・公至赤壁・与备战・不利	
	Target	公征备・与备战・不利	
	标注	十 S 二 S 月 S・S 孙 S 权 S 为 S 备 M 攻 S 合 S 肥 S・S **公 B** 自 S 江 S 陵 S 征 M 备 M・S 至 S 巴 S 丘 S・S 遣 S 张 S 憙 S 救 S 合 S 肥 S・S 权 S 闻 S 憙 S 至 S・S 乃 S 走 S・S **公 B** 至 S 赤 S 壁 S・S 与 M 备 M 战 M・S 不 S 利 E	

3.3 实验方法

(1) 模型选择

本文实验的难点在于《三国志》事件数据集较小，直接训练模型很难取得较好的成果，所以本文选取 BERT-BiLSTM-CRF-NER 模型[3]作为序列标注方法的基础模型，T5、RoBERTa、NEZHA 三个预训练模型作为文本生成方法的基础模型。

BERT（Bidirectional Encoder Representations from Transformers）[22]模型是一个基于 Transformer 结构的双向编码器，其结构可以简单理解为 Transformer 的 encoder 部分。BERT 预训练任务主要包括 MLM（Masked Language Model）和 NSP（Next Sequence Prediction）两个部分，以实现编码层的构建。该模型充分训练了包含 800M 词语的 BooksCorpus 和包含 2500M 词语的英语 Wikipedia 的大规模的无标注语料，使得下游具体任务可以很轻松的完成微调，大大降低了下游任务所需的样本数据和计算算力。BERT 模型在自然语言处理领域中的各种问题上都取得了较好的成果，其提出的预训练（pretrain）+ 微调（fine-tune）两阶段已成为自然语言处理领域的基本范式，是近几年的一大创新，随之引起了一大波预训练模型的出现。BERT 模型本文使用的 3 种文本生成模型都是在 BERT 模型的基础上进行优化改进后的预训练模型，在模型基础上结合《三国志》事件数据集微调，得到预测结果。

T5 模型（Text-To-Text-Transfer-Transformer）[23][24][25][26][27][28]是谷歌在 2019 年 10 月提出的预训练模型。从任务框架上，T5 创造性地将每个自然语言处理任务，包括自然语言理解和自然语言生成，统一为"Text-To-Text"的问题。对于机器翻译、文本分类、文本相似度、文本摘要等不同的任务，只需在输入上添加不同的前缀，即可通过生成模型得到输出结果。BERT 模型仅采用了 Transformer 架构的 encoder 部分，而 T5 模型将问题都统一成了生成问题，所以采用原版 Transformer 的 encoder 和 decoder。此外，T5 采用了相对位置编码替代固定位置编码。在数据上，T5 大大提高了训练语料的数量和质量，采用了 750GB 的 C4 语料（Colossal Clean Crawled Corpus）。在训练方法上，T5 参考 SpanBERT，采用跨度掩码（span masking）；增长训练步长，提高至 1M；使用混合训练（multi-task）的方式，在无监督数据中，加入了部分有监督的数据等等。

RoBERTa 模型（A Robusty Optimized BERT Pretraining Approach）[29]是 Facebook 与华盛顿大学在 2019 年 7 月提出的预训练模型。RoBERTa 模型相比于 BERT 模型的静态掩码，采用动态掩码，方法类似于交叉验证，这种方法使得每一份语料都会产生不同掩码，略微提高了模型的识别效果；移除了 NSP 目标函数，采用 FULL-SENTENCES 和 DOC-SENTENCES 的方式构造序列；将原始语料增大到 160G；以字节级 BPE 编码替代 BERT 模型采用的字符级 BPE 编码；超参优化，增大 batch size 和训练迭代次数等。

NEZHA（哪吒）[30]模型是华为诺亚方舟实验室在 2019 年 9 月提出的面向中文自然语言理解任务的预训练模型。在训练方法上主要作出了 4 点改进：NEZHA 模型采用函数式相对位置编码，通过使用相对位置的正弦函数计算输出和 Attention 的得分；将随机掩码替换为全词掩码进行训练；采用混合精度训练，在训练过程中同时使用单精度和半精度，从而加速训练；使用 LAMB 优化器。

(2) 集成学习

集成学习[31][32]思想由 Dasarathy 和 Sheela 在 1979 年首次提出。此后，集成学习成为机器学习领域中的一个重要分支。集成学习算法主要由 3 种经典算法组成：Bagging[33]、Boosting[34]、Stacking[35]。Bagging 算法通过 bootstrap 抽样从原始数据集生成多个不同的训练子集，再分别用不同的训练子集训练多个不同的分类器，最终采用投票的方式组合所有的分类结果。Boosting 算法通过增加迭代次数，反复运行将弱学习器转换为强学习器。Stacking 算法则是首先调用多个不同类型的个体分类器在同一训练集上进行训练，再以这些个体分类器的输出作为输入来训练元分类器。笔者参考 Stacking 算法和 Bagging 的思想，将 3 个生成模型的结果以投票的方式整合，完成生成模型的集成学习实验。

[3] https://github.com/macanv/BERT-BiLSTM-CRF-NER

3.4 模型构建

本文实验过程分为序列模型构建、生成模型构建及生成模型集成学习三个部分。

序列模型选用 BERT-BiLSTM-CRF-NER 模型作为基础模型，在模型基础上进行微调，结合《三国志》事件数据集，再训练后得到 BBCN-SG 模型，对测试集进行预测。最终，得到模型对测试集预测后的标签，所以直接对比目标标签与预测标签来评价模型性能。在应用层面，根据 BMES 标注方法，逆向将预测标签还原为预测摘要。其中，部分参数设置如下：最大序列长度 max_length 为 128，学习率设置为 1e-5，batch_size 设置为 16，训练轮数 epoch 设置为 10。

生成模型前后共选取 T5、RoBERTa、NEZHA 三个预训练模型进行实验，与序列模型相同，进行微调，结合《三国志》事件数据集再训练，构建 T5-SG、RoBERTa-SG、NEZHA-SG 模型。结合实际情况，将三个生成模型参数设置为一致，其中，最大序列长度 max_length 为 256，学习率设置为 1e-5，batch_size 设置为 16。三个生成模型的训练轮数 epoch 均设置为 20。

三个生成模型训练完成后，对测试集进行预测，得到 171*3 条预测结果。将预测结果与目标摘要文本进行对比，通过 ROUGE 与 BLEU 评估指标，对模型预测效果进行量化的展示，并在量化的基础上进行 Stacking 集成学习。

3.5 评价指标

人工评价事件识别的好坏具有较大的主观性，并且也需要耗费大量的时间和精力。因此，本文主要参考一些主流的评价指标，以量化的方式评价事件识别的结果。序列模型直接对比目标标签与预测标签，应用 seqeval 序列标注评估工具[4]进行评估。生成模型则是应用 ROUGE、BLEU 指标来评价。对于模型进行评价时，取所有预测结果得出的评价指标的平均值作为模型的评价指标。

(1) 序列模型评价指标

本文采用 BMES 的标注方式，因此直接对比目标标签与预测标签，计算准确率、召回率、查准率和 F1 值对模型性能进行评价。参考混淆矩阵，对应的计算方式如下：

$$\text{accuracy} = \frac{TP + TN}{N} \tag{1}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

其中，N 表示全部样本，TP 为实际真、预测真的样本，TN 为实际真、预测假的样本，FP 为实际假、预测真的样本，FN 为实际假、预测假的样本。

(2) BLEU 评价指标

BLEU（Bilingual Evaluation Understudy）较多用于评估机器翻译的质量，本文中通过比较生成文本与参考文本 N-gram 的重合程度来评估预测结果的好坏，两者的重合程度越高，代表预测结果越好。

BLEU 指标的计算公式如公式（5）所示，Pn 表示 N-gram 的精确率，即 N-gram 匹配的词数占总词数的比例，wn 表示 N-gram 的权重，一般取值为 1/N。由于 Pn 只针对生成文本过长的情况进行了惩罚，而没有考虑生成文本过短的情况，因此加入 BP 惩罚因子，若生成文本比参考文本长度短，就会受到简短惩罚，如公式（6）所示，其中 lc 代表生成文本的长度，lr 为参考文本的长度。

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \times \log P_n\right) \tag{5}$$

[4] seqeval: https://github.com/chakki-works/seqeval

$$BP = \begin{cases} 1 & lc > lr \\ \exp\left(1 - \frac{lr}{lc}\right) & lc \leq lr \end{cases} \quad (6)$$

历史古籍的文言文文本通常由单字或双字组成，且生成的事件长度通常较短，因此本文的 BLEU 评价指标的 N 取值为 1，2。

(3) ROUGE 评价指标

ROUGE（Recall-Oriented Understudy for Gisting Evaluation）指标由 Chin-Yew Lin [36]提出，相比于 BLEU 指标，ROUGE 更加关注召回率。ROUGE 指标共包含 ROUGE-N，ROUGE-L，ROUGE-W，ROUGE-S 四种指标。本文的 ROUGE 指标采用 ROUGE-1、ROUGE-2、ROUGE-L。

ROUGE-N 统计 N-gram 的召回率，计算公式如公式（7）所示，p 表示生成文本与参考文本中重合的 N-gram 的个数，q 表示参考文本中 N-gram 的个数。本文采用 ROUGE-N 指标的 N 取值为 1，2，即 ROUGE-1、ROUGE-2。

$$ROUGE - N = \frac{p}{q} \quad (7)$$

ROUGE-L 则考虑了生成文本与参考文本之间的最长公共子序列（Longest Common Subsequence, LCS）。ROUGE-L 计算公式（即 Flcs）如公式（8）-（10）所示，其中 C 表示生成文本，S 表示参考文本，LCS(C,S)表示 C 与 S 之间的最长公共子序列，Rlcs 代表召回率，Plcs 代表精确率，β 一般取值为很大的数值，当 β 趋近于无穷大时，Plcs 就可以忽略不计，即 Flcs 等于 Rlcs。

$$R_{lcs} = \frac{LCS(C，S)}{len(S)} \quad (8)$$

$$P_{lcs} = \frac{LCS(C，S)}{len(C)} \quad (9)$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \quad (10)$$

4 实验结果及分析

4.1 序列模型与生成模型的对比实验

序列标注方法和文本生成方法都是目前主流的事件识别方法，但学界缺少两种方法的具体比较研究。本文针对《三国志》事件数据集，从历史古籍的古汉语角度展开两种方法的对比实验。序列模型基于 BERT-BiLSTM-CRF-NER 模型，生成模型基于 T5 模型，分别进行微调实验，在原预训练模型的基础上，对《三国志》事件数据集再训练，并进行预测，两种方法的实验结果如表 4 所示。本节中，序列模型即 BBCN-SG 模型，生成模型即 T5-SG 模型。

表 4《三国志》事件识别序列模型和生成模型的评价指标得分对比

	r(%)	p(%)	f(%)
序列模型	56.37	79.62	66.61
生成模型	67.19	71.58	65.51

*r: recall　p: precision　f: F1

序列模型采用 BMES 标注方法，因此实验输出的预测结果是 BMES 标签，直接通过对比目标标签与预测标签，计算模型得分。生成模型实验输出的预测结果是文本，对比预测文本与目标文本，采用 BLEU 和 ROUGE 指标评价模型性能。表 4 展示序列模型和生成模型的召回率、准确率、F1 值，其中，生成模型选择 ROUGE-1 指标进行比较。从指标得分来看，序列模型在预测召回率上明显低于生成模型，但在准确率和 F1 值上均大于生成模型。 但本文的序列实验采取的是 BMES 标注方法，除了句首的‘B’

标签和句尾的‘E’标签，只有句中相关的‘M’标签和无关的‘S’标签，几乎可以看作是二值分类的问题，所以准确率较高并不能说明识别效果好，并且准确率对于二值分类问题来说并不高。因此，单从指标角度分析，很难区分序列模型与生成模型孰优孰劣，还需要考虑含义准确性、语义连贯性等。

由于序列模型生成的预测结果是 BMES 标签，需要按照原标签转化规则将其转化为文本，才能实现应用。但是在转化过程中存在以下的问题：按照标签文本转化规则，‘B’指目标文本中的首字，‘M’指目标文本的中间字，因此如果出现预测标签中目标文本的中间字在首字之前的情况，即‘M’在‘B’之前，很难确定其顺序，如表 5 所示，‘B’表示“吴”，第一个‘M’表示“年”，如果根据标签文本转化规则，“年”字应该在句中，但无法确认其具体的顺序，并且从语义角度出发，“年”字在句中也会影响预测文本的语义连贯性。

表 5　基于序列模型的《三国志》事件识别实例

原文本	明年·吴将陆逊向庐江
目标文本	陆逊向庐江
目标标签	['S' 'S' 'S' 'S' 'S' 'B' 'M' 'M' 'M' 'E']
预测标签	['S' 'M' 'S' 'B' 'M' 'M' 'M' 'M' 'M' 'E']
预测文本	**年**吴将陆逊向庐江

表 6 为《三国志》事件识别序列模型和生成模型的实例对比。序列模型对于远距离的约束，识别效果较差，如实例 1-3，序列模型输出的预测文本缺少语义的连贯性，难以理解，并不能有效识别历史古籍中的事件。而生成模型即使对于较长的古汉语，也能保证识别事件的语义连贯性。对于中长度的古汉语，如实例 4-7，整体上，两种模型识别的事件都能大致与原文相关，但序列模型识别事件的语义明显不如生成模型连贯，脱离原文本很难单独理解预测文本的含义，生成模型虽然部分事件的正确性不足，如实例 5，识别的事件含义与原文本不一致，但能保证整体语句可以理解。对于长度短的古汉语，两种模型均可以较好地识别事件，并能保证一定的含义准确性和语义连贯性，但生成模型识别的事件明显优于序列模型。此外，生成模型的预测文本中包含部分词组，如实例中的“长史”“合肥”“袁绍”等，而序列模型的预测文本基本还是单字为主，无法识别出一些固定词组。因此，在含义准确性和语义连贯性等角度，序列模型明显不如生成模型。

表 6　《三国志》事件识别序列模型和生成模型的实例对比

序号	原文本	目标文本	序列模型预测文本	生成模型预测文本
1	凌于后索兵不得·乃单遣一督将步骑七百人往迎之·布夜掩击·督将迸走·死伤过半	凌遣督迎布 布掩击 死伤过半	凌索不乃遣督步七往迎·夜击督迸死过	凌 擒 督 将 斩 斩 督 将
2	宠驰往赴·募壮士数十人 折松为炬·灌以麻油·从上风放火·烧贼攻具·射杀权弟子孙泰·贼于是引退	宠烧贼攻具 杀泰退贼	宠往募士十·松炬灌麻·上放·贼具射权子孙泰贼引	宠 诱 宠 杀 权
3	宠以为田向收孰·男女布野其屯卫兵去城远者数百里 可掩击也·遣长史督三军循江东下·摧破诸屯·焚烧穀物而还	宠遣长史循江东下 摧破诸屯 焚烧穀物而还	宠田向孰男布·屯兵城者百·掩也遣史军江东摧破屯 焚穀而	宠 遣 长史 督 军 破 诸 屯
4	权果上岸耀兵·宠伏军卒起击之·斩首数百·或有赴水死者	权上岸耀兵 宠伏击之 斩首数百	权岸兵宠军击斩数或赴死	权 斩 伏 军 卒
5	公孙瓒使豫守东州令·瓒将王	公孙瓒将王门叛	孙瓒守州瓒王叛瓒袁绍	袁绍 攻 袁绍

	门叛瓒·为袁绍将万余人来攻	瓒 攻东州	余来	
6	明年·权自将号十万·至合肥新城	权自将号十万 至合肥新城	年权将十至合肥城	权 自 将 号 十万 至合肥 新城
7	三年春·权遣兵数千家佃于江北	权遣兵佃于江北	年权遣千佃江	权 遣 兵 数 千 家 佃于 江北
8	明年·吴将陆逊向庐江	陆逊向庐江	年吴将陆逊向庐江	吴 将 陆 逊 向 庐江
9	豫时年少·自托于备	豫自托备	豫少自于备	豫 托 于 备
10	豫以母老求归	豫以母老求归	豫母求归	豫 求 归 豫 求 归

对于序列标注方法，事件识别不同于命名实体识别的短距离约束，而是以远距离约束为主。从预测结果来看，序列标注方法对于包含远距离约束的事件并不能较好的识别，预测结果的含义、语义等均较差。从预测结果的平均指标得分来看，序列模型的召回率低于生成模型，但准确率和 F1 值略高于生成模型。但从预测结果具体的含义准确性和语义连贯性来看，序列模型远不如生成模型。综上，序列模型在《三国志》事件识别上的表现不如生成模型。

4.2 生成模型的对比实验

对比序列模型和生成模型在《三国志》事件识别上的表现，可以发现生成模型具有更好的识别效果，因此增加 RoBERTa 和 NEZHA 两个预训练模型，在《三国志》事件数据集上再微调训练，对比三个生成模型的识别效果，以 BLEU 和 ROUGE 指标评价模型性能，具体得分对比如表 7 所示。

表 7　三个《三国志》事件识别生成模型的 BLEU 及 ROUGE 评价指标得分对比

（%）	BLEU-1	BLEU-2	ROUGE-1			ROUGE-2			ROUGE-L		
			r	p	f	r	p	f	r	p	f
T5-SG	47.04	33.46	**67.19**	71.58	65.51	**48.58**	45.17	44.04	**64.70**	68.65	63.02
RoBERTa-SG	**51.75**	**38.78**	63.49	**75.77**	**65.70**	45.52	**53.29**	**46.35**	62.07	**73.48**	**64.07**
NEZHA-SG	46.86	32.13	57.32	71.77	60.83	37.37	42.49	37.76	55.93	69.91	59.35

* r: recall　p: precision　f: F1

BLEU 指标注重精确率，从 BLEU 指标来看，RoBERTa-SG 模型在三个模型中得分最高，T5-SG 模型次之，但与 NEZHA-SG 模型相差不大。ROUGE 指标的精确率同样如此，RoBERTa ＞ T5-SG≈NEZHA-SG。ROUGE 指标更注重召回率，ROUGE-1、ROUGE-2、ROUGE-L 指标的召回率均是 T5-SG 模型得分最高，RoBERTa-SG 模型次之，NEZHA-SG 模型最低。因此，如图 2 所示，三个模型从召回率角度，T5-SG 模型最好；从精确率角度，RoBERTa-SG 模型最好，NEZHA-SG 模型表现不如其他两个模型。仅从指标得分来看，将所有指标相加，RoBERTa-SG 在 3 个生成模型中得分最高，3 个模型的表现依次是 RoBERTa-SG ＞ T5-SG ＞ NEZHA-SG。

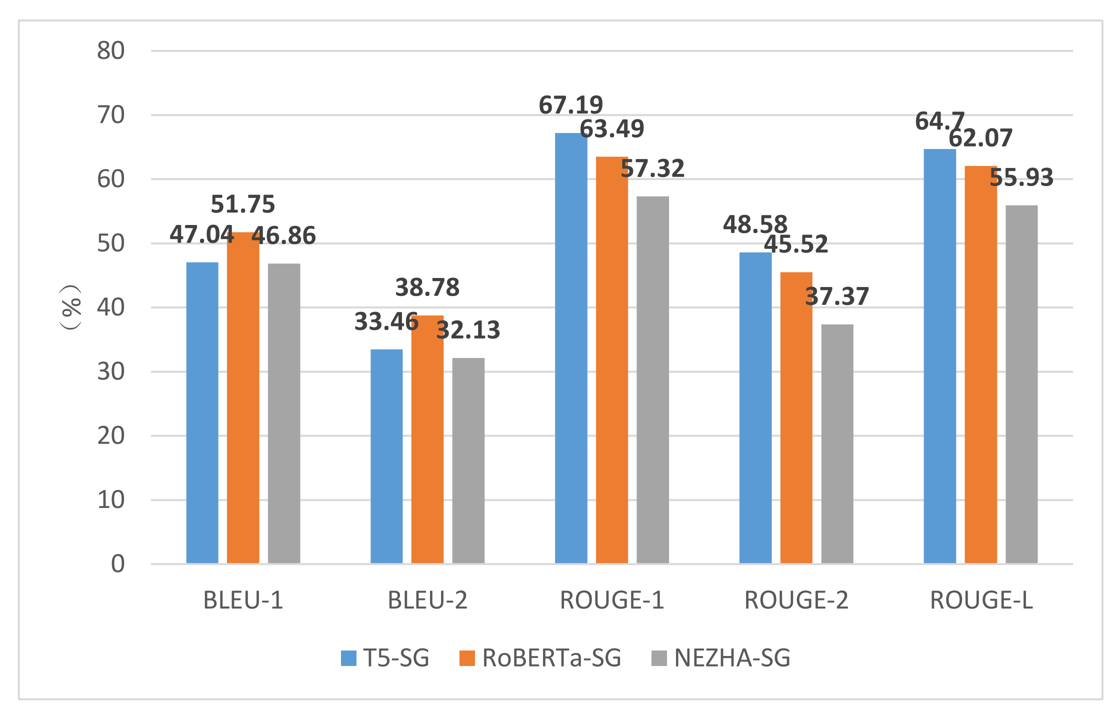


图 2 三个《三国志》事件识别生成模型的 BLEU 及 ROUGE 评价指标对比
（ROUGE 指标选取召回率进行对比）

4.3 生成模型的集成学习实验

虽然 RoBERTa-SG 模型在《三国志》事件识别上的整体表现最好，但从召回率的角度来看，T5-SG 模型得分最高。因此，笔者进一步探究 3 个生成模型在具体实例上的表现。

如表 8 所示，在具体实例上，3 个生成模型并非完全是某一种模型的表现最好，而是存在一定的互补性。以 BLEU-2 指标为衡量标准，实例 1 是 NEZHA-SG 模型得分最高，实例 2 是 RoBERTa-SG 模型得分最高，实例 3 是 T5-SG 模型得分最高，实例 4 是三个模型得分相同。因此，联想到 Stacking 集成学习的思想，将 3 个生成模型结合，可以大大提高模型的识别效果。Stacking 集成学习思想可以简单理解为并行学习多个异质弱学习器，并通过一个“元”模型将它们组合起来，输出一个最终的预测结果。由于 3 个生成模型的模型结构并不相同，很难将 3 个生成模型直接组合，结合 Bagging 集成学习的方法，采取投票的方式聚合预测结果。以 BLEU 指标为投票标准，对三个生成模型的预测结果进行排序，最终以得分最高的预测结果作为 Stacking 集成学习的最终结果。据 BLEU 指标的原理，BLEU-2 指标比 BLEU-1 指标的要求更为苛刻，因此①先以 BLEU-2 指标排序；②若 BLEU-2 指标相同，则以 BLEU-1 指标排序；③若 BLEU-2、BLEU-1 指标均相同，则以 ROUGE 指标排序，由于 ROUGE 指标注重召回率，而 T5-SG 模型的召回率最高，所以按照 T5-SG > RoBERTa-SG > NEZHA-SG 的顺序选择预测结果为最终结果。表 8 中①列表示考虑 BLEU-2 指标后的排序结果，②列表示综合考虑 BLEU-2、BLEU-1 指标后的排序结果，③列表示最终的排序结果，即 Stacking 集成学习的最终结果。

表 8　3 个《三国志》事件识别生成模型具体指标对比

序号	T5	RoBERTa	NEZHA	①	T5	RoBERTa	NEZHA	②	③
	BLEU-2				BLEU-1				
1	0.00%	0.00%	9.10%	NEZHA	15.16%	4.93%	24.56%	NEZHA	NEZHA
2	0.00%	3.51%	0.00%	RoBERTa	12.11%	10.54%	13.38%	RoBERTa	RoBERTa
3	100.00%	80.43%	30.16%	T5	100.00%	90.48%	54.29%	T5	T5
4	66.67%	66.67%	66.67%	相同	71.43%	71.43%	71.43%	相同	T5
5	23.08%	50.00%	50.00%	Ro\NE	28.57%	57.14%	57.14%	Ro\NE	RoBERTa

6	0.00%	0.00%	0.00%	相同	10.74%	10.74%	17.97%	NEZHA	NEZHA
7	0.00%	0.00%	0.00%	相同	16.67%	8.33%	0.00%	T5	T5
8	72.73%	72.73%	60.00%	T5\Ro	75.00%	75.00%	72.73%	T5\Ro	T5
9	22.47%	9.55%	22.47%	T5\NE	35.95%	21.49%	44.93%	NEZHA	NEZHA
10	81.87%	44.44%	81.87%	T5\NE	81.87%	50.00%	81.87%	T5\NE	T5

将集成学习后的预测文本与目标文本进行对比，计算每一条文本的 BLEU 和 ROUGE 指标得分，求均值作为 Stacking-TRN-SG 模型的最终得分。得分如表 9 所示，所有指标均高于 3 个基础的生成模型，说明模型的识别效果取得了较大的提高。

表 9　Stacking-TRN-SG 模型 BLEU 及 ROUGE 评价指标得分

（%）	BLEU-1	BLEU-2	ROUGE-1			ROUGE-2			ROUGE-L		
			r	p	f	r	P	f	r	p	f
Stacking-TRN-SG	**63.40**	**51.04**	**70.35**	**77.72**	**71.42**	**54.45**	**57.25**	**53.61**	**68.78**	**75.90**	**69.83**
T5-SG	47.04	33.46	67.19	71.58	65.51	48.58	45.17	44.04	64.70	68.65	63.02
RoBERTa-SG	51.75	38.78	63.49	75.77	65.70	45.52	53.29	46.35	62.07	73.48	64.07
NEZHA-SG	46.86	32.13	57.32	71.77	60.83	37.37	42.49	37.76	55.93	69.91	59.35

具体集成学习生成的实例如表 10 所示。集成学习后的生成模型取三者最优，识别结果的含义更加准确、语义也更为连贯。无论是长文本还是短文本，Stacking-TRN-SG 输出的识别事件都容易理解、长度也很合适，能够保留古汉语单字居多的特点。然而，部分含义仍然不够准确，还需要进一步研究。

表 10　3 个《三国志》事件识别生成模型的识别实例对比

序号	原文本	目标文本	T5-SG	RoBERTa-SG	NEZHA-SG	Stacking-TRN-SG
1	凌于后索兵不得 乃单遣一督将步骑七百人往迎之 布夜掩击督将迸走 死伤过半	凌遣督迎布 布掩击 死伤过半	凌 擒 督 将 斩 斩 督 将	凌 单 兵 不 得	凌 遣 骑 骑 千 人 往 破 督 督	凌 遣 骑 骑 千 人 往 破 督 督
2	宠驰往赴 募壮士数十人 折松为炬 灌以麻油 从上风放火 烧贼攻具 射杀权弟子孙泰 贼于是引退	宠烧贼攻具 杀泰退贼	宠 诱 宠 杀 权	宠 引 火 烧 贼	宠 遣 人 壮 赶 赶 贼	宠 引 火 烧 贼
3	宠以为田向收孰 男女布野 其屯卫兵去城远者数百里 可掩击也 遣长史督三军循江东下 摧破诸屯 焚烧穀物而还	宠遣长史循江东下 摧破诸屯 焚烧穀物而还	宠 遣 长史 督 军 破 诸 屯	宠 向 田 向 破 诸 屯	宠 遣 三 军 破 之 屯	宠 遣 长史 督 军 破 诸 屯
4	权果上岸耀兵 宠伏军卒起击之 斩首数百 或有赴水死者	权上岸耀兵 宠伏击之 斩首数百	权 斩 伏 军 卒	权 斩 兵 宠	权 斩 斩 军 斩 斩 斩 卒	权 斩 兵 宠
5	公孙瓒使豫守东州令 瓒将王门叛瓒 为袁绍将万余人来攻	公孙瓒将王门叛瓒 攻东州	袁绍 攻 袁 绍	瓒 使 王 门 叛 瓒	公 孙 瓒 叛 瓒	瓒 使 王 门 叛 瓒
6	明年 权自将号十万 至合肥新城	权自将号十万 至合肥新城	权 自 将 号 十万 至 合肥 新城	权 将 号 十 万 至 合 肥 新 城	权 率 率 五 万 进 合 肥 新 城	权 自 将 号 十 万 至 合肥 新 城
7	三年春 权遣兵数千家佃于江北	权遣兵佃于江北	权 遣 兵 数 千 家 佃 于 江北	权 遣 兵 数 千 家 于 江 北	权 遣 兵 佃 于 江 北	权 遣 兵 佃 于 江 北
8	明年 吴将陆逊向庐江	陆逊向庐江	吴 将 陆 逊 向 庐江	吴 将 陆 逊 向 庐 江	吴 将 陆 逊 向 庐 江	吴 将 陆 逊 向 庐江
9	豫时年少 自托于备	豫自托备	豫 托 于 备	豫 时 少 托 于 备	豫 少 少 自 托 备	豫 少 少 自 托 备
10	豫以母老求归	豫以母老求归	豫 求 归 豫 求 归	豫 以 母 老 求 归	豫 以 老 老 求 归	豫 以 母 老 求 归

5 总结

历史古籍是传承中华文化的重要载体，基于历史古籍构建知识图谱可以直观地向人们展示历史，而事件识别是构建知识图谱的重要一环，因此实现历史古籍的自动事件识别是很有必要的。由于古汉语具有单字居多、语义难以理解等特点，学界缺少对于古汉语的事件识别研究。此外，序列标注方法和文本生成方法是目前主流的两种事件识别方法，但缺少对于两种方法的具体实例对比研究。因此，本文选取《三国志》为原始语料，分别从序列标注与文本生成两个方法展开实验。序列标注实验对《三国志》事件数据集进行 BMES 标注，应用 BERT-BiLSTM-CRF-NER 模型进行训练和预测。文本生成实验则应用 T5 预训练模型进行训练和预测。选取 BLEU 与 ROUGE 评价指标对预测结果量化评价，并结合含义准确性、语义连贯性等因素评估模型性能。实验发现，文本生成方法在《三国志》事件数据集上的表现大大超过序列标注方法。又选取 RoBERTa、NEZHA 两个预训练模型进行事件识别训练。从预测结果上来看，三个模型的整体表现则是 RoBERTa-SG > T5-SG > NEZHA-SG，但也存在一定的互补性。融

入 Stacking 集成学习的思想，结合三个文本生成模型，构建 Stacking-TRN-SG 模型，相比于 3 个基础的生成模型，识别效果大大提高。Stacking-TRN-SG 模型识别出的事件能够保证一定的含义准确、语义连贯，并且可以体现出古汉语单字居多等特点，召回率也达到 70.35%，取得了较好的识别效果，初步实现历史古籍的自动事件识别。

参考文献

[1] 鄂海红,张文静,肖思琪,程瑞,胡莺夕,周筱松,牛佩晴.深度学习实体关系抽取研究综述[J].软件学报,2019,30(06):1793-1818.DOI:10.13328/j.cnki.jos.005817.

[2] 赵妍妍,秦兵,车万翔,刘挺.中文事件抽取技术研究[J].中文信息学报,2008(01):3-8.

[3] 姜吉发.自由文本的信息抽取模式获取的研究[D]．中国科学院博士学位论文，2004:1-18.

[4] Mihai Surdeanu, Sanda Harabagiu, John Williams, etal. Using Predicate-Argument Structures for Information Extraction [A]．In :Proceedings of ACL[C]．2003.8-15.

[5] Mihai Surdeanu, Sanda Harabag iu ．Infrastructurefor Open-Domain Information Extraction [A]．In:Proceedings of the Human Language Technology Conference [C]．2002.325-330.

[6] Zhou H, Chen J, Dong G, et al. Detection and Diagnosis of Bearing Faults Using Shift-invariant Dictionary Learning and Hidden Markov Model[J]. Mechanical Systems & Signal Processing, 2015(72-73):65-79.

[7] 卢达威， 宋柔． 基于最大熵模型的汉语标点句缺失话题自动识别初探[J]． 计算机工程与科学， 2015， 37(12):2282-2293.

[8] 李丽双，黄德根，毛婷婷，等．基于支持向量机的中国人名的自动识别[J]．计算机工程，2006，32(19)：188-190.

[9] 邬伦， 刘磊， 李浩然，等． 基于条件随机场的中文地名识别方法[J]． 武汉大学学报:信息科学版， 2017，42(2):150-156.

[10] Ahn D. The stages of event extraction[C]//Proceedings of the COLING-ACL 2006 Workshop on Annotating and Reasoning About Time and Eyents, 2006: 1—8.

[11] 王子牛,姜猛,高建瓴,陈娅先.基于 BERT 的中文命名实体识别方法[J].计算机科学,2019,46(S2):138-142.

[12] 任智慧,徐浩煜,封松林,周晗,施俊.基于 LSTM 网络的序列标注中文分词法[J].计算机应用研究,2017,34(05):1321-1324+1341.

[13] 陈伟,吴友政,陈文亮,张民.基于 BiLSTM-CRF 的关键词自动抽取[J].计算机科学,2018,45(S1):91-96+113.

[14] 唐慧慧,王昊,张紫玄,王雪颖.基于汉字标注的中文历史事件名抽取研究[J].数据分析与知识发现,2018,2(07):89-100.

[15] Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1724-1734.

[16] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 3104-3112.

[17] Rush A M, Chopra S, Weston J. A neural attention model for abstractive sentence summarization[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015:379-389.

[18] 石磊,阮选敏,魏瑞斌,成颖.基于序列到序列模型的生成式文本摘要研究综述[J].情报学报,2019,38(10):1102-1116.

[19] 刘忠宝,党建飞,张志剑.《史记》历史事件自动抽取与事理图谱构建研究[J].图书情报工作,2020,64(11):116-124.DOI:10.13266/j.issn.0252-3116.2020.11.013.

[20] 李章超，李忠凯，何琳．《左传》战争事件抽取技术研究[J].图书情报工作,2020,64(07):20-

29.DOI:10.13266/j.issn.0252-3116.2020.07.003.
[21] 喻雪寒,何琳,徐健.基于RoBERTa-CRF的古文历史事件抽取方法研究[J].数据分析与知识发现,2021,5(07):26-35.
[22] Devlin J, Chang M W, Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// 2018.
[23] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. J. Mach. Learn. Res., 2020, 21(140): 1-67.
[24] Xue L, Constant N, Roberts A, et al. mT5: A massively multilingual pre-trained text-to-text transformer[J]. arXiv preprint arXiv:2010.11934, 2020.
[25] Shazeer N. Glu variants improve transformer[J]. arXiv preprint arXiv:2002.05202, 2020.
[26] Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks[C]//International conference on machine learning. PMLR, 2017: 933-941.
[27] Chung H W, Fevry T, Tsai H, et al. Rethinking embedding coupling in pre-trained language models[J]. arXiv preprint arXiv:2010.12821, 2020.
[28] Joshi M, Chen D, Liu Y, et al. Spanbert: Improving pre-training by representing and predicting spans[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 64-77.
[29] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
[30] Wei J, Ren X, Li X, et al. Nezha: Neural contextualized representation for chinese language understanding[J]. arXiv preprint arXiv:1909.00204, 2019.
[31] 徐继伟,杨云.集成学习方法:研究综述[J].云南大学学报(自然科学版),2018,40(06):1082-1092.
[32] 周星,丁立新,万润泽,葛强.分类器集成算法研究[J].武汉大学学报(理学版),2015,61(06):503-508.DOI:10.14188/j.1671-8836.2015.06.001.
[33] Breiman L. Bagging predictors[J]. Machine Learning, 1996,24(2):123-140.
[34] Freund Y. Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences,1997,55(1):119-139
[35] Wolpert D H. Stacked generalization[J]. Neural networks, 1992:241-259.
[36] Lin C Y . ROUGE: A Package for Automatic Evaluation of summaries[C]// In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). 2004.

（通讯作者：王昊，E-mail：ywhaowang@ nju.edu.cn）

作者贡献说明

王彦莹：完成实验、论文起草、撰写与修改；
王昊：确定研究思路、指导实验、提出论文框架、指导论文修改；
朱惠：指导论文修改；
李晓敏：指导论文修改。